GenAI-Powered Inference

Kosuke Imai

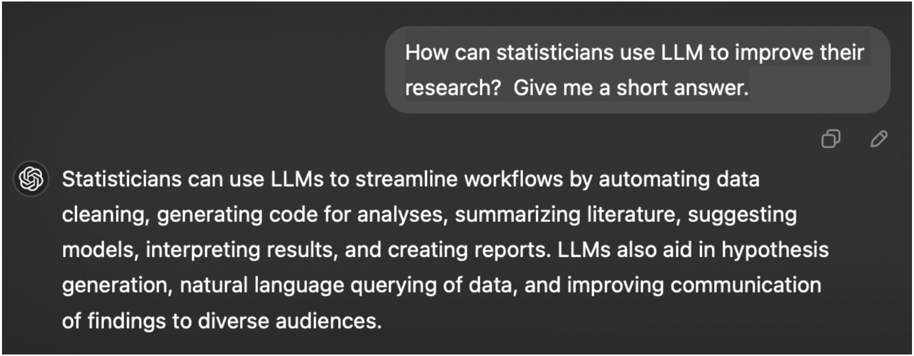
Harvard University

Techstars Tokyo
October 9, 2025

Joint work with Kentaro Nakamura (Harvard Kennedy School)

Motivation

- Generative AI is transforming medicine, education, marketing, etc.
- Can statisticians get some help from GenAI too?



Using GenAI to Improve Statistical Inference

- GenAI-Powered Inference (GPI)
- GenAI-assisted causal/predictive inference with unstructured data
  1. (re)generate unstructured data at scale
  2. obtain true internal representation from GenAI
  3. use it directly for machine learning without fine tuning

- Advantages:
  - no need to estimate representation
  - avoid functional form assumptions
  - better empirical performance

Generative AI: Definition and Assumption

- Deep generative model:

$$\mathbb{P}(\boldsymbol{X}_i \mid \boldsymbol{h}_\gamma(\boldsymbol{R}_i)),$$
$$\mathbb{P}(\boldsymbol{R}_i \mid \boldsymbol{P}_i).$$

  - $\boldsymbol{P}_i$: prompt
  - $\boldsymbol{X}_i$: unstructured object generated by GenAI
  - $\boldsymbol{R}_i$: hidden states or internal representations
  - $\boldsymbol{h}_\gamma(\boldsymbol{R}_i)$: deterministic function from hidden states to the last layer

- Deterministic decoding:

$$\mathbb{P}(\boldsymbol{X}_i \mid \boldsymbol{h}_\gamma(\boldsymbol{R}_i)) \text{ is degenerate}$$

  - can be achieved by setting a hyperparameter
  - use of open-source GenAI and deterministic encoding for replicability

Motivating Application: Texts-as-Treatments

- Candidate Biography Experiment (Fong and Grimmer, 2016)
    - 1246 biographies of American politicians scraped from Wikipedia
    - 1,886 voters as respondents
    - randomly assign biographies to voters
    - feeling thermometer $[0, 100]$ as the outcome

- Analysis
    - supervised topic model to discover 10 treatment features
    - estimate the average treatment effects of estimated topic proportions

- Existing approaches for texts-as-treatments:
    1. model-based approach (e.g., Egami *et al.* 2022; Fong and Grimmer, 2023)
    2. causal representation learning based on fine-tuned BERT embedding (e.g., Veitch et al. 2020; Pryzant et al. 2021; Gui and Veitch, 2023)

Example Biographies

Candidate biography with military background
Anthony Higgins was born in Red Lion Hundred in New Castle County, Delaware. He attended Newark Academy and Delaware College, and graduated from Yale College in 1861, where he was a member of Skull and Bones. After studying law at the Harvard Law School, he was admitted to the bar in 1864 and began practice in Wilmington, Delaware. He also served for a time in the United States Army in 1864.
Candidate biography without military background
Benjamin Tappan was born in Northampton, Massachusetts, the second child and oldest son of Benjamin Tappan and Sarah (Homes) Tappan, who was a grandniece of Benjamin Franklin. Two of his younger brothers were abolitionists Arthur Tappan and Lewis Tappan. He attended the public schools in Northampton and traveled to the West Indies in his youth. He apprenticed as a printer and engraver, also studying painting with Gilbert Stuart. He read law to be admitted to the bar in Hartford, Connecticut, in 1799. Later that year, he moved to the Connecticut Western Reserve and founded what is now Ravenna, Ohio, laying out the original village in 1808. He married, March 20, 1801, Nancy Wright, sister of John C. Wright (congressman), afterwards a United States House of Representatives from Ohio. They had one son, Benjamin, born in 1812.

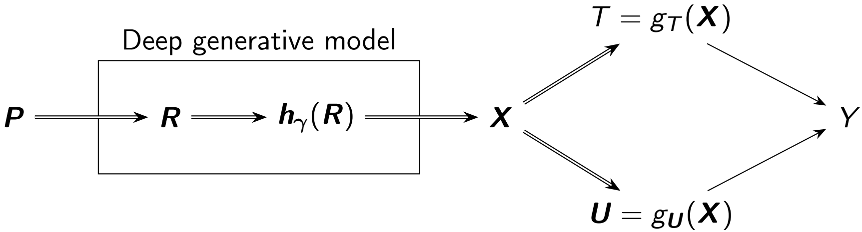
Setup

- Notation
  - $Y_i(\boldsymbol{x})$: Potential outcome when exposed to treatment object $\boldsymbol{x}$
  - $Y_i = Y_i(\boldsymbol{X}_i)$: Outcome (collected from the survey respondents)
  - T_i: Binary treatment feature (e.g., military experiences)
  - $\boldsymbol{U}_i$: Confounding features (e.g., college education)

- Assumptions
  1. Treatment Feature:
  $$T_i \;=\; g_T(\boldsymbol{X}_i)$$
  2. Confounding Features:
  $$\boldsymbol{U}_i \;=\; \boldsymbol{g_U}(\boldsymbol{X}_i) \quad \text{where } \dim(\boldsymbol{U}_i) \ll \dim(\boldsymbol{X}_i)$$
  3. Separability:
  $$Y_i(\boldsymbol{x}) = Y_i(g_T(\boldsymbol{x}), \boldsymbol{g_U}(\boldsymbol{x})),$$
     1. T is not a function of $\boldsymbol{U}$
     2. $\boldsymbol{U}$ is not a function of T

Summary of Assumptions

- Assumptions:



- Overlap: The above assumptions imply that for any $t \in \{0, 1\}$ and $\boldsymbol{u} \in \mathcal{U}$, we have

$$\mathbb{P}(T_i = t \mid \boldsymbol{U}_i = \boldsymbol{u}) > 0.$$

Nonparametric Identification

- Average treatment effect (ATE):

$$\tau \ := \mathbb{E}[Y_i(1, \boldsymbol{U}_i) - Y_i(0, \boldsymbol{U}_i)]$$

- Under these assumptions, there exists a **Deconfounder** $\boldsymbol{f} : \mathbb{R}^r \to \mathbb{R}^q$ with $q \leq r$ such that

$$Y_i \perp\!\!\!\perp \boldsymbol{R}_i \mid T_i = t, \boldsymbol{f}(\boldsymbol{R}_i), \quad t \in \{0, 1\}$$

  - Deconfounder does not have to be unique
  - Example: Confounding Features $\boldsymbol{U}_i$ (deterministic function of $\boldsymbol{R}_i$)

- By adjusting for this Deconfounder, we can identify the marginal distribution of potential outcome as

$$\mathbb{P}(Y_i(t, \boldsymbol{U}_i) = y) \ = \ \int_{\mathbb{R}^r} \mathbb{P}(Y_i = y \mid T_i = t, \boldsymbol{f}(\boldsymbol{R}_i)) dF(\boldsymbol{R}_i),$$

- Direct adjustment for $\boldsymbol{R}_i$ leads to the lack of overlap

Estimation and Inference

$$
\begin{tikzcd}
 & & \mu_1(\boldsymbol{f}(\boldsymbol{R}_i;\boldsymbol{\lambda});\boldsymbol{\theta}_1) \arrow[r] & Y_i \mid T_i = 1\text{: Treated obs.} \\
\boldsymbol{R}_i \arrow[r] & \boldsymbol{f}(\boldsymbol{R}_i;\boldsymbol{\lambda}) \arrow[ur] \arrow[dr] & & \\
 & & \mu_0(\boldsymbol{f}(\boldsymbol{R}_i;\boldsymbol{\lambda});\boldsymbol{\theta}_0) \arrow[r] & Y_i \mid T_i = 0\text{: Control obs.}
\end{tikzcd}
$$

1. Estimate the outcome models and deconfounder via TarNet (Shalit et al. 2017):

$$\{\hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\theta}}_0, \hat{\boldsymbol{\theta}}_1\} \;=\; \underset{\boldsymbol{\lambda},\boldsymbol{\theta}_0,\boldsymbol{\theta}_1}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n}\{Y_i - \mu_{T_i}(\boldsymbol{f}(\boldsymbol{R}_i;\boldsymbol{\lambda});\boldsymbol{\theta}_{T_i})\}^2$$

2. Estimate the propensity score using the estimated Deconfounder

$$\pi(\boldsymbol{f}(\boldsymbol{R}_i, \hat{\boldsymbol{\lambda}})) = \mathbb{P}(T_i = 1 \mid \boldsymbol{f}(\boldsymbol{R}_i, \hat{\boldsymbol{\lambda}}))$$

Popular DragonNet (Shi et al. 2019) jointly estimates the outcome models, propensity score, and deconfounder, leading to the lack of overlap

Double Machine Learning (Chernozhukov et al. 2018)

- Cross-fitting:
  1. randomly divide the data into K folds
  2. for each $k = 1, \ldots, K$, use the kth fold as the test set and the remaining $k-1$ folds as the training set
     1. randomly split the training set further into two subsets
     2. use the first subset to estimate outcome models and deconfounder
     3. use the second subset to estimate propensity score given the estimated deconfounder
  3. Compute the ATE estimator as:

$$
\begin{aligned}
\hat{\tau} \;=\; & \frac{1}{nK} \sum_{k=1}^{K} \sum_{i:I(i)=k} \hat{\mu}_1^{(-k)}(\hat{\boldsymbol{f}}^{(-k)}(\boldsymbol{R}_i)) - \hat{\mu}_0^{(-k)}(\hat{\boldsymbol{f}}^{(-k)}(\boldsymbol{R}_i)) \\
& + \frac{T_i\{Y_i - \hat{\mu}_1^{(-k)}(\hat{\boldsymbol{f}}^{(-k)}(\boldsymbol{R}_i))\}}{\hat{\pi}^{(-k)}(\hat{\boldsymbol{f}}^{(-k)}(\boldsymbol{R}_i))} - \frac{(1-T_i)\{Y_i - \hat{\mu}_0^{(-k)}(\hat{\boldsymbol{f}}^{(-k)}(\boldsymbol{R}_i))\}}{1-\hat{\pi}^{(-k)}(\hat{\boldsymbol{f}}^{(-k)}(\boldsymbol{R}_i))}
\end{aligned}
$$

- Double robustness, asymptotic normality

Practical Implementation Details

- Internal representation extracted from LLM is still high-dimensional:

  $$\dim(\boldsymbol{R}) = \text{number of tokens} \times 4096 \quad \text{for Llama 3 (8 billion parameters)}$$

- Pooling strategies depend on deep generative models
  - BERT: the first special classification token [CLS]
  - Llama 3: the hidden states of the last token

- TarNet requires hyperparameter tuning
  - size and depth of layers
  - learning rate
  - maximum epoch size
- Use of automatic hyperparameter optimization methods (e.g., Optuna)

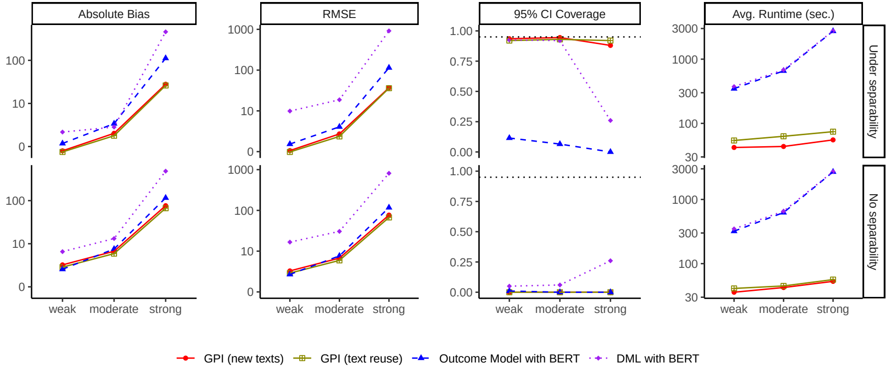
Simulation Study Setup

- A simulation based on the candidate biography experiment
  - Create 4,000 sets of the first, middle, and last names of political candidates via randomly sampling from the Fong and Grimmer data
  - Use Llama 3 to generate a biography for each US political candidate's
  - Instruct LLM to repeat the same texts for reuse
- The data generating process:

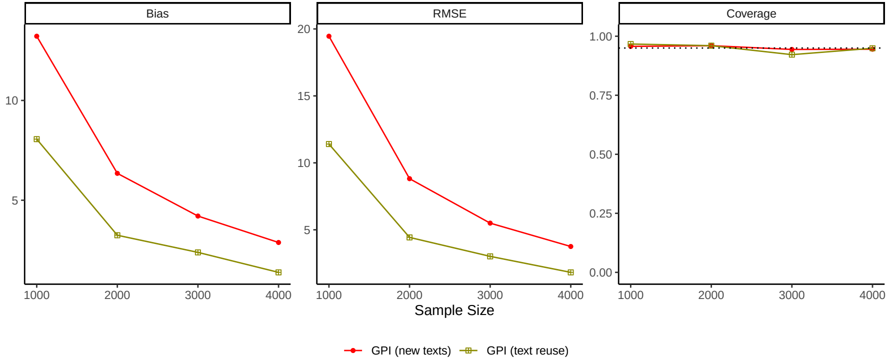
$$Y_i = \alpha_1 T_i + \alpha_2 T_i h_1(\boldsymbol{X}_i) - \alpha_3 h_1(\boldsymbol{X}_i) - \alpha_4 h_2(\boldsymbol{X}_i) + \epsilon_i$$

  where $\epsilon_i \sim \mathcal{N}(\mu_i, 1)$ and
  - T_i: military background (binary)
  - $h_1(\boldsymbol{X}_i)$: topic-model based confounder
  - $h_2(\boldsymbol{X}_i)$: sentiment-analysis based confounder
- $2 \times 3 = 6$ scenarios:
  1. separability holds or does not hold (separate or overlapping topics)
  2. weak, medium, or strong confounding

Simulation Results



Performance across Different Sample Sizes



Empirical Analysis

- Analyze the original survey by Fong and Grimmer (2016)
  - 1,246 Congressional candidate biographies from Wikipedia
  - 1,886 survey participants with a total of 5,291 observations
  - evaluate a biography using the feeling thermometer [0, 100]
  - Keyword-based treatment coding: "military", "war", "veteran", or "army"
  - use text-reuse approach with Llama 3

Methods	ATE	95% Conf. Int.	Runtime
GPI (reuse)	5.462	[2.790, 8.135]	28.9 sec.
Outcome model with BERT	−2.557	[−2.608, −2.505]	6139.7
DML with BERT	−67.777	[−109.967, −25.587]	6210.3

Predictive Effects of Image Features (Lindholm et al. 2024)

- How does the visual appearance of political candidate predict their electoral success?
- Data: 7,080 Danish politicians with candidate photos

- Prediction variables: facial features (continuous scores)
  1. attractiveness
  2. trustworthiness
  3. dominance
- Discretize them into 10 bins
- Outcome: Election results (number of votes standardized via z-score)
- Structured confounding variables: age, gender, education
- We wish to adjust other facial confounding features

Empirical Analysis

- Reproduce all images using Stable diffusion (ver. 1.5; also version 2.1)
- Original image:

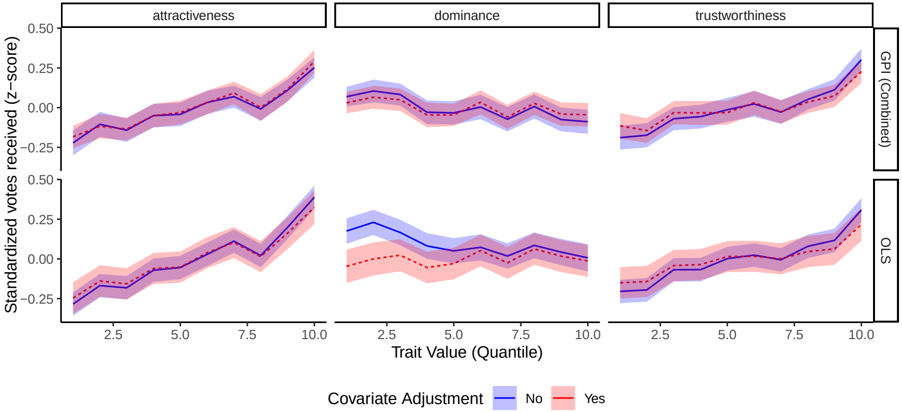
$$\dim(\boldsymbol{X}) = 304(\text{width}) \times 304(\text{height}) \times 3(\text{RGB}) = 277248$$

- Internal representation: $\dim(\boldsymbol{R}) = 16384$
- Neural network architecture:
  - $\dim(\boldsymbol{f}(\boldsymbol{R})) = 1024$
  - depth of hidden layers $= 2$
  - size of hidden layers after deconfounder $= [200, 1]$
- Nonparametrically estimate the average predictive effect

$$\xi_t := \mathbb{E}[Y_i(t, \boldsymbol{U}_i)]$$

- DML with a multi-valued treatment

Empirical Results



- Unlike OLS, the proposed method is not sensitive to the inclusion of structured confounding variables

Structural Model of Texts: Persuasion and Rhetoric

(Blumenau and Lauderdale, 2022)

- Which types of political rhetorics are most persuasive?

- Forced choice conjoint experiment with texts
- Total of 336 political arguments
  - 12 policy issues: tuition fees, fracking, etc.
  - 14 rhetorical elements: cost and benefit, morality, etc.
  - for or against

- Outcome: Persuasiveness of arguments
  - one argument is more persuasive than the other
  - equally persuasive

Example Text Pair

- Policy topic: building a third runaway at Heathrow:

Appeal to authority / For The Airports Commission, an independent body established to study the issue, have argued that expanding Heathrow is the most effective option to address the UK's aviation capacity challenge
Appeal to history / Against History show us that most large infrastructure projects do not lead to significant economic growth, which suggests that the expansion of Heathrow will fail to pay for itself

- Can we adjust for the unstructured confounding features of texts?

The Structural Model

- The original Bradley-Terry type model:

$$\log\left[\frac{\mathbb{P}(Y_{jj'(i)} \leq k)}{\mathbb{P}(Y_{jj'(i)} > k)}\right] = \delta_k + \left(\alpha_{P_j S_j} + \beta_{T_j} + \gamma_j\right) - \left(\alpha_{P_{j'} S_{j'}} + \beta_{T_{j'}} + \gamma_{j'}\right)$$

  where i indexes respondents, j indexes arguments, P_j denotes policy area, S_j denotes for/against, and T_j denotes rhetoric

- Our semiparametric model:

$$\log\left[\frac{\mathbb{P}(Y_{j(i),j'(i)} \leq k)}{\mathbb{P}(Y_{j(i),j'(i)} > k)}\right] = \delta_k + \mu(T_j, \boldsymbol{U}_j) - \mu(T_{j'}, \boldsymbol{U}_{j'})$$

- Persuasiveness of rhetoric $T_j = t$

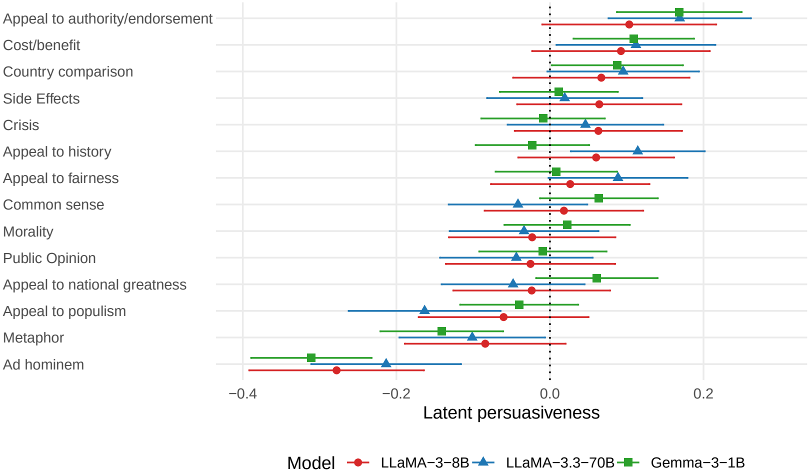
$$\beta(t) := \mathbb{E}[\mu(t, \boldsymbol{U}_j)]$$

- Estimate $\beta(t)$ using the deconfounder $\boldsymbol{f}(\boldsymbol{R}_j)$

Empirical Analysis

- Reproduce all texts using Llama3–8B (also Llama3.3–70B and Gemma3.1–1B)
- Internal representation: last token of the final layer, $\dim(\boldsymbol{R}) = 4096$
- Neural network architecture:
  - $\dim(\boldsymbol{f}(\boldsymbol{R})) = 1024$
  - depth of hidden layers $= 2$
  - size of hidden layers after deconfounder $= [200, 1]$
- Quantify uncertainty via Monte Carlo dropout (Gal and Ghahramani 2016)

Empirical Findings



- Stronger effects for ad hominem, appeal to authority, and cost/benefit
- Findings are similar across models
- Smaller standard errors for newer models

Concluding Remarks

- Generative AI can be used to improve causal inference
  - can generate treatments at scale
  - enables the extraction of true internal representation
  - better causal representation learning

- Open-source software GPI is available at

  https://gpi-pack.github.io/

- Further extensions
  - causal inference with multimodal data (e.g., videos)
  - interpretation of estimated deconfounder
  - discovery of treatment concepts
  - policy learning with unstructured treatments